Methoden-Workshop

Leading House „Economics of Education“/SKBF

Ben Jann und Rudi Farys

Universität Zürich, Rämistrasse 71, Raum KOL-F-123, 1.–3. Februar 2016

Hauptkomponentenanalyse und Faktorenanalyse

Hauptkomponentenanalyse vs. Faktorenanalyse

- Hauptkomponentenanalyse (Principal Component Analysis, PCA)
  - Stata: `pca`, `pcamat`
  - Datenreduktionsverfahren: Reduktion einer Menge von Variablen auf eine kleinere Anzahl an Komponenten (Linearkombinationen der Variablen), die orthogonal (unkorreliert) sind und einen möglichst grossen Anteil der Varianz der Originalvariablen abbilden.
- (Explorative) Faktorenanalyse
  - Stata: `factor`, `factormat`
  - Exploratives Verfahren zur Analyse von Messinstrumenten: Auffinden von „latenten“ Faktoren („Konstrukten“), die hinter der „gemeinsamen Varianz“ (i.e. der Kovarianzstruktur) einer Menge von Messvariablen stehen.
- Konfirmatorische Faktorenanalyse
  - Stata: `sem`, `gsem`
  - Überprüfung von Messmodellen: Inwieweit stimmt die beobachtete Datenstruktur mit dem unterstellten Modell „latenter“ Konstrukte, die sich auf die Messvariablen auswirken, überein.

Hauptkomponentenanalyse vs. Faktorenanalyse

- Hauptkomponentenanalyse und (explorative) Faktorenanalyse haben einen unterschiedlichen konzeptionellen Hintergrund, in der praktischen Anwendung gibt es aber viele Parallelen.
  - Konzeptionell entgegengesetzte Betrachtungsweise, die jedoch letztlich zu qualitativ ähnlichen Resultaten führt:
    - Hauptkomponentenanalyse: Die Komponenten bzw. Faktoren werden als Linearkombinationen der Messvariablen ausgedrückt.
    - Faktorenanalyse: Die Messvariablen werden als lineare Funktion der Faktoren (plus Fehlerterm) ausgedrückt.
- Hauptkomponentenanalyse und (explorative) Faktorenanalyse sind beides Verfahren zur Datenexploration. Es geht darum (vereinfachende) Strukturen in einer Menge von Messvariablen zu finden.
- Die konfirmatorische Faktorenanalyse funktioniert umgekehrt. Hier wird die Struktur der Messmodells vorgegeben. Es wird dann geprüft, ob die beobachteten Daten dazu passen.

Hauptkomponentenanalyse

- p (standardisierte) Variablen werden zu p „Komponenten“ transformiert.
- Bei den „Komponenten“ handelt es sich um orthogonale (i.e. unkorrelierte) Linearkombinationen der p Variablen.
- Die Gewichte der Linearkombinationen werden als „Ladungen“ bezeichnet. Die Ladungen sind so normiert, dass die Summe der quadrierten Ladungen für jede Komponente gleich eins ist.
- Das Ziel ist, mit möglichst wenigen Komponenten einen möglichst grossen Anteil der Varianz der (standardisierten) Ausgangsvariablen zu erklären.
  - Die erste Komponente ist diejenige Linearkombination, die den grösstmöglichen Anteil der Gesamtvarianz abbildet.
  - Die zweite Komponente ist diejenige Linearkombination, die den grösstmöglichen Anteil der verbleibenden Varianz abbildet, die nicht in der ersten Komponente enthalten ist.
  - etc.

Hauptkomponentenanalyse

- Die Komponenten werden durch sog. Eigen-Dekomposition der Korrelationsmatrize der Variablen bestimmt. Die „Eigenwerte“ entsprechen den Varianzen der Komponenten, die „Eigenvektoren“ enthalten die Ladungen.
  - Die Summe der Eigenwerte ist gleich p, da standardisierte Variablen eine Varianz von 1 haben (i.e. die Summe der Varianzen von p standardisierten Variablen ist gleich p).
  - Die erste Komponente hat den höchsten Eigenwert, die zweite den zweithöchsten, etc.
- I.d.R. werden nur k Komponenten mit den grössten Eigenwerten weiter betrachtet.
  - z.B. nur Komponenten mit Eigenwert > 1; bis zum Knick im Scree-Plot; mindestens X% der Varianz
- Voraussetzung für eine sinnvolle Anwendung der Hauptkomponentenanalyse: Intervallskalenniveau (Linearkombinationen machen sonst keinen Sinn).

Hauptkomponentenanalyse

- Beispiel: Messung von Umweltbewusstsein (http://www.farys.org/daten/uba98.dta)

v11a	Es beunruhigt mich, wenn ich daran denke, unter welchen Umweltverhältnissen unsere Kinder und Enkelkinder wahrscheinlich leben müssen.
v11b	Wenn ich Zeitungsberichte über Umweltprobleme lese oder entsprechende Fernsehmeldungen sehe, bin ich oft empört und wütend.
v11c	Wenn wir so weitermachen wie bisher, steuern wir auf eine Umweltkatastrophe zu.
v11d	Nach meiner Einschätzung wird das Umweltproblem in seiner Bedeutung von vielen Umweltschützern stark übertrieben. (umgekehrt gepolt)
v11e	Es ist noch immer so, dass die Politiker viel zu wenig für den Umweltschutz tun.
v10e	Umweltschutzmassnahmen sollten auch dann durchgesetzt werden, wenn dadurch Arbeitsplätze verloren gehen.
v11f	Egal, was die anderen tun, ich selbst versuche, mich soweit wie mäglich umweltgerecht zu verhalten.
v11g	Zugunsten der Umwelt sollten wir alle bereit sein, unseren Lebensstandard einzuschränken.
v11h	Ich verhalte mich auch dann umweltbewusst, wenn es zusätzlich erheblich höhere Kosten und Mühen verursacht.

Hauptkomponentenanalyse: Rotation

- Die Komponenten sind häufig schwierig zu interpretieren, deshalb werden sie meistens noch rotiert.
- Meistens wird orthogonal rotiert, so dass auch die rotierten Komponenten unkorreliert sind.
- Das am häufigsten verwendete Rotationsverfahren ist die sog. Varimax-Rotation. Die Rotation erfolgt dabei so, dass die Varianz der Ladungen innerhalb der Komponenten maximiert wird. Das führt tendenziell dazu, dass die Variablen auf die Komponenten „aufgeteilt“ werden.
- Durch die Rotation ändern sich die Eigenwerte der Komponenten, die Summe der Eigenwerte der rotierten Komponenten bleibt jedoch erhalten.
- Rotation kann aber auch wenig zielführend sein. Siehe z.B. das Beispiel mit den Hörtests in [MV] **pca** bzw. [MV] **pca postestimation**

(Explorative) Faktorenanalyse

- Bei der Faktorenanalyse werden p beobachtete Variablen als Funktion von q gemeinsamen Faktoren ausgedrückt:

$$x_{ij} = b_{1j}z_{i1} + b_{2j}z_{i2} + \cdots + b_{2j}z_{i2} + e_{ij}, \quad j = 1, \ldots, p$$

  (z sind die Faktoren; b die Ladungen; e ist der Rest, der nicht durch die gemeinsamen Faktoren erklärt wird)
- Die Resultate werden zur einfacheren Interpretation üblicherweise ebenfalls rotiert.
- Neben dem Varimax-Verfahren gibt es noch eine Vielzahl weiterer Rotationsverfahren. Je nach dem werden auch nicht-orthogonale Rotationen verwendet. Insgesamt ist das allerdings alles eine ziemliche Kaffeesatzleserei.